(Client Ref. No. P01-1654)

UNITED STATES PATENT APPLICATION

FOR

Reconfiguration of Storage System Including Multiple Mass Storage Devices

INVENTORS:

Brad A. Reger
Susan M. Coatney

Prepared by:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP
12400 WILSHIRE BOULEVARD
SEVENTH FLOOR
LOS ANGELES, CALIFORNIA 90025
(408) 720-8300

Attorney's Docket No. 5693P037X

"Express Mail" mailing label number EV409364896US
Date of Deposit April 15, 2004
I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to Mail Stop Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

Julie Arango
(Typed or printed name of person mailing paper or fee)

Julie Arango 4/15/04
(Signature of person mailing paper or fee)

Reconfiguration of Storage System Including Multiple Mass Storage Devices

[0001] This application is a continuation-in-part of:

[0002] U.S. Patent application no. 10/027,457 of S. Coatney et al., filed on December 21, 2001 and entitled, "System and Method of Implementing Disk Ownership in Networked Storage" (hereinafter "Coatney");

[0003] U.S. Patent application no. 10/027,020 of J. Sen Sarma et al., filed on December 21, 2001 and entitled, "System and Method for Transferring Volume Ownership in Networked Storage" (hereinafter "Sarma");

[0004] U.S. Patent application no. 10/027,013 of A. Rowe et al., filed on December 21, 2001 and entitled, "System and Method for Allocating Spare Disks in Networked Storage" (hereinafter "Rowe"); and

[0005] U.S. Patent application no. 10/407,681 of B. Reger et al., filed on April 4, 2003 and entitled, "Method and Apparatus for Converting Disk Drive Storage Enclosure into a Standalone Network Storage System and Vice Versa" (hereinafter "Reger");

each of which is incorporated herein by reference.

FIELD OF THE INVENTION

[0006] At least one embodiment of the present invention pertains to data storage systems, and more particularly, to a technique for reconfiguring a storage system.

BACKGROUND

[0007] Modern computer networks can include various types of storage servers. Storage servers can be used for many different purposes, such as to provide multiple users with access to shared data or to back up mission critical data. A file server is one type of storage server, which operates on behalf of one or more clients to store and

manage shared files in a set of mass storage devices, such as magnetic or optical storage based disks or tapes. The mass storage devices are typically organized into one or more volumes of Redundant Array of Independent (or Inexpensive) Disks (RAID).

[0008] One configuration in which a file server can be used is a network attached storage (NAS) configuration. In a NAS configuration, a file server can be implemented in the form of an appliance, called a filer, that attaches to a network, such as a local area network (LAN) or a corporate intranet. An example of such an appliance is any of the Filer products made by Network Appliance, Inc. in Sunnyvale, California.

[0009] A storage server can also be employed in a storage area network (SAN). A SAN is a highly efficient network of interconnected, shared storage devices. In a SAN, the storage server (which may be an appliance) provides a remote host with block-level access to stored data, whereas in a NAS configuration, the storage server provides clients with file-level access to stored data. Some storage servers, such as certain Filers from Network Appliance, Inc. are capable of operating in either a NAS mode or a SAN mode, or even both modes at the same time. Such dual-use devices are sometimes referred to as "unified storage" devices. A storage server such as this may use any of various protocols to store and provide data, such as Network File System (NFS), Common Internet File system (CIFS), Internet SCSI (ISCSI), and/or Fibre Channel Protocol (FCP).

[0010] Historically, file server systems used in NAS environments have generally been packaged in either of two forms: 1) an all-in-one custom-designed system that is essentially just a standard computer with built-in disk drives, all in a single chassis; or 2) a modular system in which one or more sets of disk drives (each set being mounted in a

separate chassis) are connected to a separate external file server "head". Examples of all-in-one file server systems are the F8x, C1xxx and C2xxx series Filers made by Network Appliance, Inc. Examples of modular filer heads are the F8xx and FAS9xx heads made by Network Appliance, Inc.

[0011] In this context, the term "head" means all of the electronics, firmware and/or software that is used to control access to storage devices in a storage system; it does not include the disk drives themselves. In a file server, the head normally is where all of the "intelligence" of the file server resides. Note that a "head" in this context is not the same as, and is not to be confused with, the magnetic or optical head used to physically read or write data to a disk.

[0012] In a modular file server system, the system can be built up by adding multiple disk enclosures in some form of rack and then cabling the disk enclosures together. The disk drive enclosures are often called "shelves", and more specifically, "just a bunch of disks" (JBOD) shelves. The term JBOD indicates that the enclosure essentially contains only physical storage devices and no substantial electronic "intelligence". Some disk drive enclosures include one or more RAID controllers, but such enclosures are not normally referred to as "JBOD" due to their greater functional capabilities.

[0013] Modular storage systems and all-in-one storage systems each have various shortcomings, as noted in Reger (referenced above). Reger describes a standalone network storage server that overcomes some of the shortcomings of modular storage systems and all-in-one storage systems. The standalone storage server includes multiple internal single-board heads and multiple internal disk drives, all contained within a single chassis and connected to each other by an internal passive backplane.

Each head contains the electronics, firmware and software along with built-in I/O connections to allow the disks in the enclosure to be used as a NAS file server and/or a SAN storage device.

[0014] Reger also describes that the standalone storage server can be easily converted into a JBOD shelf (essentially, by removing the internal heads and replacing them with I/O modules) and then integrated into a modular storage system such as described above. This allows a storage system to be grown in capacity and/or performance by combining the converted JBOD shelf with one or more separate (modular), more-powerful file server heads, such as Network Appliance F8xx or FAS9xx series heads, and additional JBOD shelves.

[0015] Although this convertability makes the standalone storage server very versatile, reconfiguring a storage system in this manner is not a trivial task. This type of system reconfiguration can require fairly extensive rerouting and addition of cables to allow the modular heads to control the disks in the newly-converted JBOD shelf (converted from the standalone storage server). In many storage systems with redundant heads, each disk is "owned" by (primarily accessed by) only one head, and disk ownership is determined by the cable connections. For example, in some systems, each disk has two external ports, port A and port B, which are connected (at least indirectly) to two separate heads. Only the head connected to port A owns the disk, while the head connected to port B assumes a backup role for purposes of accessing that disk.

[0016] To integrate a converted JBOD shelf into a modular system in the manner described above requires reassigning ownership of all of the disks in the converted JBOD shelf (which were owned by the removed internal heads) to an external modular

head. As indicated above, such reassignment of ownership can require moving disks from one enclosure to another as well as extensive rerouting of cables and/or addition of new external cabling, all of which is inconvenient and complicated.

SUMMARY OF THE INVENTION

[0017] The present invention includes a method in which a storage system, which includes a plurality of mass storage devices and a first storage server head to access the mass storage devices in response to client requests, is operated, wherein the first storage server head has ownership of the plurality of mass storage devices. Ownership of at least one of the mass storage devices is reassigned to a second storage server head, independently of a manner in which the second storage server head is connected to the plurality of mass storage devices.

[0018] Other aspects of the invention will be apparent from the accompanying figures and from the detailed description which follows.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] One or more embodiments of the present invention are illustrated by way of example and not limitation in the figures of the accompanying drawings, in which like references indicate similar elements and in which:

[0020] Figure 1 illustrates a storage system that includes a storage server head and a set of JBOD shelves in a “rack and stack” configuration;

[0021] Figure 2 is a block diagram showing how a file server head can be connected to a set of clients and a set of JBOD shelves;

[0022] Figure 3 is a block diagram similar to that of Figure 2 but illustrating one of the JBOD shelves in greater detail;

[0023] Figure 4 shows a standalone storage server which includes two single-board heads and a set of disks;

[0024] Figure 5 shows a storage system which includes the standalone storage server coupled to a set of expansion JBOD shelves;

[0025] Figure 6 is a flow diagram illustrating a process for converting the standalone storage server into a JBOD shelf and then integrating it, along with its expansion JBOD shelves, into a more powerful modular storage system; and

[0026] Figures 7 through 16 show the storage system in various states throughout the process of Figure 6.

DETAILED DESCRIPTION

[0027] A method and apparatus to integrate a JBOD shelf, which has been converted from a standalone storage server, into a modular storage system are described. Note that in this description, references to “one embodiment” or “an embodiment” mean that the feature being referred to is included in at least one embodiment of the present invention. Further, separate references to “one embodiment” or “an embodiment” in this description do not necessarily refer to the same embodiment; however, such embodiments are also not mutually exclusive unless so stated, and except as will be readily apparent to those skilled in the art from the description. For example, a feature, structure, act, etc. described in one embodiment may also be included in other embodiments. Thus, the present invention can include a variety of combinations and/or integrations of the embodiments described herein.

[0028] As mentioned above, Reger describes a standalone network storage server that overcomes some of the shortcomings of modular and all-in-one storage systems. Reger also describes how the standalone storage server can be easily converted into a JBOD shelf and then integrated into a modular storage system. Described herein is a technique to integrate a JBOD shelf which has been converted from a standalone storage server, such as described in Reger, into a modular storage system.

[0029] Figure 1 illustrates an example of a modular file server system arranged in a “rack and stack” configuration. In Figure 1, a file server head 1 is connected by external cables to multiple disk shelves 2 mounted in a rack 3. The file server head 1 enables access to stored data by one or more remote client computers (not shown) that are connected to the head 1 by external cables. The modular file server head 1 may be, for example, a F8xx or FAS9xx series Filer made by Network Appliance.

[0030] Figure 2 is a functional block diagram of a modular file server system such as shown in Figure 1. The modular file server head 1 is contained within its own enclosure and is connected to a number of the external JBOD shelves 2 in a (logical) loop configuration. Each JBOD shelf 2 contains multiple disk drives 23 operated under control of the head 1 according to RAID protocols. The file server head 1 provides a number of clients 24 with access to shared files stored in the disk drives 23. Note that Figure 2 shows a simple network configuration characterized by a single loop with three shelves 2 in it; however, other network configurations are possible. For example, there can be a greater or smaller number of JBOD shelves 2 in the loop; there can be more than one loop attached to the head 1; or, there can even be one loop for every JBOD shelf 2.

[0031] Figure 3 illustrates an example of a JBOD shelf 2 in greater detail (clients 24 are not shown). Each of the shelves 2 can be assumed to have the same construction. Each shelf 2 includes multiple disk drives 23. Each shelf also includes at least one I/O module 31, which is connected between the shelf 2 and the next shelf 2 in the loop and in some cases (depending on where the shelf 2 is placed in the loop) to the head 1. The I/O module 31 is a communications interface between the head 1 and the disk drives 23 in the shelf 2. The disk drives 23 in the shelf 2 can be connected to the I/O module 31 by a standard Fibre Channel connection, for example.

[0032] The I/O module 31, in addition to acting as a communications interface between the head 1 and the disk drives 23, also serves to enhance reliability by providing loop resiliency. Thus, in certain embodiments each I/O module 31 is a Loop Resiliency Circuit (LRC). If a particular disk drive 23 within a shelf 2 is removed or fails, the I/O module 31 in that shelf 2 simply bypasses the missing or failed disk drive and

connects to the next disk drive within the shelf 2. In certain embodiments this functionality maintains connectivity of the loop in the presence of disk drive removals and is provided by multiple Port Bypass Circuits (PBCs) (not shown) included within the I/O module 31 (typically, a separate PBC for each disk drive 23 in the shelf 2).

[0033] Figure 4 is a hardware layout block diagram of a standalone storage server such as described in Reger. The standalone storage server 71 includes multiple disk drives 23, multiple heads 64, and a passive backplane 51, all of which are contained within a single chassis. Each of the heads 64 is implemented on a separate, single circuit board. An example of the architecture of the single-board head 64 is described in Reger. The heads 64 and disk drives 23 are all connected to, and communicate via, a passive backplane 51. The storage server 71 further includes a power supply 52 and a cooling module 53 for each head 64.

[0034] The standalone storage server 71 can be easily grown in capacity and/or performance by combining it with additional modular JBOD shelves 2, as shown in Figure 5, and (optionally) with one or more separate, more powerful file server heads. Alternatively, the standalone storage server 71 can be converted into a JBOD shelf 2 by removing and replacing each of the heads 64 with an I/O module, such as I/O module 31 described above. The JBOD shelf thus created can then be integrated into a more powerful, modular storage system of the type described above regarding Figures 1 through 3. This allows the standalone storage system 71 to be easily upgraded by the user into a more powerful storage system.

[0035] A process of integrating a JBOD shelf, converted in this way, into a modular storage system will now be described with reference to Figures 5 through 16. For purposes of description, it is assumed that the starting point for the conversion is a

storage system 50 shown in Figure 5, which comprises the standalone storage server 71 coupled to two expansion JBOD shelves 2 in a daisy chain physical topology (which logically may form a loop such as shown in Figures 2 and 3). The standalone storage server 71 will be converted into a JBOD shelf and then integrated with a more powerful storage system 120, shown in Figure 12. The more powerful system 120 initially comprises two modular (separate, external) storage heads 121, which may be, for example, FAS9xx series heads from Network Appliance, coupled to two expansion JBOD shelves 2 in a daisy chain physical topology.

[0036] It is further assumed that the single-board heads 64 in the standalone storage server 71, as well as the modular heads 121 (see Figure 12) to which the converted shelf will be integrated, all support software (command) based assignment and modification of disk ownership, in the manner described in Coatney and Sarma (referenced above). As described in detail in Coatney and Sarma, disk ownership can be determined by storing disk ownership information (including the identity of the head which owns the disk) in a predetermined area on each disk. In this way, disk ownership can be assigned independently of the manner in which the head is connected to the disks, i.e., independently of the cabling configuration between the disks and the heads. This approach contrasts with prior techniques in which disk ownership was determined entirely by the cabling configuration between the disks and the heads. Note that the conversion process could be carried out in a system which does not implement disk ownership in this manner; however, the process would be more complicated, since it would require more extensive recabling between devices to implement the desired disk ownership scheme. Finally, it is assumed that the single-board heads 64 and the

modular heads 121 all support the commands which are described as being input to them in the process which follows.

[0037] Figure 6 illustrates a process for converting the standalone storage server 71 into a JBOD shelf 2 and then integrating it, along with its expansion JBOD shelves 2, into a more powerful modular storage system 120 such as shown in Figure 12. Initially, system 50 contains the standalone storage server 71 and two expansion JBOD shelves 2, as shown in Figure 5. It is assumed that the more powerful, modular system 120, into which it will be integrated, initially uses topology base disk ownership, i.e., disk ownership is determined by the physical connections between disks and heads. To begin the conversion process, therefore, the modular system 120 is first converted to a software-based ownership scheme, such as described in Coatney and Swarma.

[0038] The process begins at block 602, in which a network administrator inputs a "disk show" command to each modular head (602), which produces a display that identifies all disks owned by the modular heads 121. This command and the other commands described below may be input from an administrative console (not shown) that is connected to the modular heads 121 either directly or over a network. If all disks that are physically connected to the modular heads 121 are indicated as being assigned to a modular head, then the process proceeds to block 604. Otherwise, the administrator inputs a "disk upgrade ownership" command to each modular head 121, which causes all disks connected to each modular head 121 to be assigned to the modular head to which it is connected (and an indication of ownership to be stored on each disk) in the manner described in Coatney (603.1). The administrator then inputs the "disk show" command again to verify that all disks in the modular system have been properly assigned (603.2).

[0039] At block 604, the network administrator inputs a "halt" command to each of the modular heads 121 (Figure 12). Next, the administrator inputs a “halt” command (block 605) to each of the single-board heads 64 in the standalone storage server 71 (Figure 5). This may be done in essentially the same manner as for the modular heads 121, i.e., through an administrative console connected (either directly or indirectly) to the single-board heads. The halt command has the effect of flushing all user data to disks.

[0040] Next, the administrator powers down the standalone storage server 71 (block 606) along with its expansion JBOD shelves 2 and then disconnects the standalone storage server's network connections (not shown) (block 607). The administrator then removes each of the single-board heads 64 from the standalone storage system 71 (block 608), as shown in Figures 7 and 8. Next, as shown in Figures 9 and 10, the administrator installs I/O modules 31 in place of the removed single-board heads 64 (block 609). At this point, the unit which contained the standalone storage server 71 is no longer a standalone storage server, but is instead a JBOD shelf 2, as shown in Figure 11. To distinguish this newly created JBOD shelf from the other JBOD shelves 2, this device is henceforth referred to as JBOD shelf 2’.

[0041] Figure 12 shows the two modular heads 121, which will control the system once the conversion process is complete. The new system 120 initially comprises the two modular heads and two expansion JBOD shelves 2, identified as New Shelf #1 and New Shelf #2. The old system 50, which includes the newly created JBOD shelf 2' and its corresponding expansion shelves 2, will be integrated with the new system 120.

[0042] To continue to process, the administrator next changes the shelf identifiers (IDs) of the expansion shelves 2 in the old system 50 as necessary to make them unique with respect to the expansion shelves 2 in the new system 120 (block 610). It

may be assumed that the shelf ID can be set by a physical switch on each shelf. For example, as shown in Figure 13, Old Shelf #1, Old Shelf #2 and Old Shelf #3 in the old system 50 are renamed as Old Shelf #3, Old Shelf #4 and Old Shelf #5, respectively. The administrator then connects cables appropriately (block 611) to connect the shelves 2 and 2' of the old system 50 to the new system 120, as shown in Figure 14. The administrator then appropriately configures the system to reflect the desired closed loop shelf topology (block 612). This may be accomplished using any of various techniques, such as by appropriately setting loop termination switches on the I/O modules, connecting a loopback plug to the downstream port of a shelf, etc.

[0043] Next, the administrator adds additional shelves to the modular heads 121, if desired (block 613), as shown in Figure 15, and then connects the cluster interconnect between the modular heads 121 and connects the modular heads 121 to the network (not shown) (block 614), as shown in Figure 16. The administrator then powers on and boots up the modular heads 121 (block 615). Figure 16 shows the final physical configuration of the system (excluding the clients and the network).

[0044] At this point, none of the disks from the old system 50 have a valid owner, since the single-board heads 64 are gone. Therefore, ownership of those disks must be reassigned. Continuing the process, therefore, the administrator next inputs a "disk show" command to modular head #1 (block 616) from an appropriate administrative console, to obtain a display identifying all of the disks in the system and an indication of the owner of each disk. The resulting display identifies all of the disks originally in the new system 120 as well as all of the disks from the old system 50. The resulting display indicates, however, that the disks from the old system 50 are not currently owned by modular head #1. Accordingly, the administrator inputs a "disk reassign"

command to modular head #1 (block 617), passing as a parameter the name of single-board head #1. This command causes the disks previously owned by single- board head #1 to be reassigned to modular head #1. That is, modular head #1 now owns those disks as a result of this command. Examples of the specific actions performed in response to such a command to change disk ownership are described in Coatney and Sarma.

[0045] Next, the administrator inputs the "disk show" command to modular head #2 (block 618) from an appropriate administrative console. As indicated above, this command produces a display identifying all of the disks from the old system 50 and the new system 120. The administrator then inputs the "disk reassign" command to modular head #2 (block 619), passing as a parameter the name of single-board head #2. This command causes the disks previously owned by single-board head #2 to be reassigned to modular head #2.

[0046] The administrator then inputs the "disk show" command to display all disks in the system and verifies that all disks are owned by the correct head (block 620). Next, the administrator uses the "disk assign" command to assign any unowned drives (block 621) (some drives may be unowned if they were added to the system in block 613). Assuming ownership is verified to be correct, it is still necessary to reassign ownership at the volume level. A storage system such as described herein may comprise multiple "volumes", each of which may comprise multiple disks. After reassignment of disk ownership is complete, the volume(s) formed by the disks from the old system 50 will appear as foreign volumes to the modular heads 121. To correct this condition, therefore, the administrator inputs a "volume on-line" command to all of the modular heads 121, to reassign the volume(s) formed by the disks from the old system 50 to the

modular heads 121 (block 622). If desired, the administrator can also delete the old root volume at this point, after which the conversion process is complete.

[0047] Thus, a method and apparatus to integrate a JBOD shelf, which has been converted from a standalone storage server, into a modular storage system have been described. Although the present invention has been described with reference to specific exemplary embodiments, it will be recognized that the invention is not limited to the embodiments described, but can be practiced with modification and alteration within the spirit and scope of the appended claims. Accordingly, the specification and drawings are to be regarded in an illustrative sense rather than a restrictive sense.